A General and Robust Framework for Secondary Traits Analysis

Xiaoyu Song,^{*,1} Iuliana Ionita-Laza,[†] Mengling Liu,[‡] Joan Reibman,[§] and Ying Wei[†]

^{*}Heilbrunn Department of Population and Family Health and [†]Department of Biostatistics, Columbia University, New York, New York 10032, and [‡]Department of Population Health and [§]Department of Medicine, New York University School of Medicine, New York, New York 10016

ORCID ID: 0000-0003-1909-6244 (X.S.)

ABSTRACT Case-control designs are commonly employed in genetic association studies. In addition to the case-control status, data on secondary traits are often collected. Directly regressing secondary traits on genetic variants from a case-control sample often leads to biased estimation. Several statistical methods have been proposed to address this issue. The inverse probability weighting (IPW) approach and the semiparametric maximum-likelihood (SPML) approach are the most commonly used. A new weighted estimating equation (WEE) approach is proposed to provide unbiased estimation of genetic associations with secondary traits, by combining observed and counterfactual outcomes. Compared to the existing approaches, WEE is more robust against biased sampling and disease model misspecification. We conducted simulations to evaluate the performance of the WEE under various models and sampling schemes. The WEE demonstrated robustness in all scenarios investigated, had appropriate type I error, and was as powerful or more powerful than the IPW and SPML approaches. We applied the WEE to an asthma case-control study to estimate the associations between the thymic stromal lymphopoietin gene and two secondary traits: overweight status and serum IgE level. The WEE identified two SNPs associated with overweight in logistic regression, three SNPs associated with serum IgE levels in linear regression, and an additional four SNPs that were missed in linear regression to be associated with the 75th quantile of IgE in quantile regression. The WEE approach provides a general and robust secondary analysis framework, which complements the existing approaches and should serve as a valuable tool for identifying new associations with secondary traits.

KEYWORDS secondary trait analysis; estimating equations; case-control studies

GENOME-WIDE association studies (GWAS) have been widely used to detect the association between common genetic variants and complex traits (Visscher *et al.* 2012) and are commonly conducted using case-control designs. In addition to the primary binary trait used to define the case-control status, data on secondary traits are often collected. For example, in a chronic obstructive pulmonary disease study (Regan *et al.* 2010), researchers collected information on additional respiratory diseases, such as asthma, emphysema, and bronchitis. It is of interest to take full advantage of the existing data and analyze genetic associations with these additional traits. Such secondary analyses have great

potential to discover additional variants associated with the secondary traits.

Several classical methods for the direct analysis of secondary traits are available, including direct regression in (1) a combined sample of cases and controls, (2) cases only, (3) controls only, and (4) combined cases and controls, adjusting for the primary disease. These methods, although attractive due to the simplicity of model building, can be biased when the secondary phenotype is associated with the primary disease. This is because the case-control sample is no longer a representative sample of the general population. Numerous studies have demonstrated that none of these methods result in an unbiased estimation of the genetic association with the secondary trait such as in Jiang *et al.* (2006), Richardson *et al.* (2007), and Monsees *et al.* (2009).

More sophisticated methods aiming to avoid the aforementioned bias can be broadly classified into two groups. The first group is based on the inverse probability weighting (IPW) idea

that is widely used in survey methodology. Jiang *et al.* (2006), Richardson *et al.* (2007), and Monsees *et al.* (2009) introduced the IPW approach to genetic studies when case and control sampling probabilities are available. The main idea of IPW is to reweight individual observations by the inverse of their selection probabilities and conduct weighted regressions. This approach requires the information of the case-control sampling scheme to construct the weights and is often biased or inefficient when there are unobserved confounders and overweighted subsamples.

The second group explicitly accounts for the case-control sampling scheme by maximizing the retrospective likelihood function conditional on the primary disease or joint modeling of the primary and secondary traits (Jiang *et al.* 2006; Lin and Zeng 2009; He *et al.* 2011; Wang and Shete 2011, 2012; Li and Gail 2012; Ghosh *et al.* 2013; Wei *et al.* 2013). The semi-parametric maximum likelihood (SPML) proposed by Lin and Zeng (2009) is the most widely recognized approach in this group as it largely improves the efficiency over IPW. This approach makes the linear logit assumption for the probability of primary disease with respect to genotypic score and secondary trait. It is very sensitive to its assumption and can be largely biased when it is violated. Li and Gail (2012) proposed an adaptive weighted approach aiming at improving the robustness of SPML by weighting the estimates from SPML and its extension (including the interactive effects of genetic markers and secondary trait in predicting case-control status in its assumption). However, the efficiency was lost and simulations showed that its biases can be as large as 46% (Wang and Shete 2012). He *et al.* (2011), Wang and Shete (2011, 2012), Ghosh *et al.* (2013), and Wei *et al.* (2013) approached this issue from different perspectives, but they were all based on a similar model specification of primary disease, and their estimation efficiency over SPML was marginal. Therefore, for the comparison presented in this article, we focus our work mainly on the IPW and SPML approaches.

Despite the tremendous efforts on secondary analysis, there are a number of remaining issues. First, the performance of existing methods heavily depends on either the sampling scheme or the correct specification of the primary disease model. Since many case-control studies of complex diseases do not have clear selection probabilities and distribution of primary disease, the estimations from these existing methods can be biased. Second, for the novel approaches that are likelihood based, additional challenges arise for regressions with no parametric-likelihood functions. For example, the FTO genotype is associated not only with the mean but also with the variance of body mass index (BMI) (Frayling *et al.* 2007). Instead of considering only the mean of phenotypes in GWAS, it is valuable to systematically examine how the markers influence the location, scale, and shape of the entire trait distribution through quantile regression (Koenker and Bassett 1978). Such quantile regressions can identify additional markers missed by mean regressions that are associated with certain quantiles of the trait. Existing secondary analysis approaches are not developed to facilitate these

types of regressions, and the expansion from likelihood-based secondary approaches is especially difficult.

In this article, we propose a set of weighted estimating equation (WEE) approaches, providing unbiased estimation for the secondary analysis in genetic case-control studies. We introduce a new concept of counterfactual estimating functions under alternative disease status and combine the observed and counterfactual estimating functions into a set of weighted estimating equations. The counterfactual outcomes idea has been widely used in causal inference to denote the potential outcomes of the subjects if they were in the alternative treatment or exposure group, and conclusions are made if the actual and counterfactual outcomes from the same subjects differ. Here, we borrow this idea to define the potential secondary trait of the subjects if they were in the alternative case-control status and estimate the marker-secondary trait association, using the case-control sample. In comparison with the existing approaches, it provides a very generalizable and robust regression framework for analyzing secondary phenotypes with limited information on the sampling scheme and the underlying primary disease models. It is flexible to handle various types of secondary phenotypes regardless of their relationship with the primary disease. After first considering the WEE for quantile regression in Wei *et al.* (2015), we expand this idea to a more general framework that accommodates a wide range of regressions. This work is outlined in this article, which is structured as follows. First, we introduce the construction of the WEE approach. Second, we present the simulations that investigate the finite sample performance under different scenarios. Finally, we consider a real data example of an asthma case-control study with one binary secondary trait (overweight status) and one continuous secondary trait [serum immunoglobulin E (IgE) levels] to illustrate our approach.

Proposed Method

Notations and settings

Let X denote the coded genotype for a variant of interest, Y denote a secondary phenotype, $\mathbf{Z}$ denote the vector of covariates to adjust for, and $D = \{0, 1\}$ denote the primary disease status. The aim of the secondary trait analysis is to estimate the genetic effect of X onto Y in the general population. The relation between Y and X and $\mathbf{Z}$ can be modeled as

$$g(Y) = \beta_0 + X\beta_1 + \mathbf{Z}^T\boldsymbol{\beta}_2, \quad (1)$$

where $g(\cdot)$ is a link function, and β_1 is the coefficient of primary interest. Depending on the choice of g , model (1) covers a wide range of regressions. For example, if $g(Y) = E[Y|X, \mathbf{Z}]$ for continuous Y , then it is linear regression; if $g(Y) = \text{logit}\{P(E[Y] = 1|X, \mathbf{Z})\}$ for binary Y , then it is logistic regression; if $g_\tau(Y) = Q_Y(\tau|X, \mathbf{Z})$ is the quantile function at the τ th quantile for Y , then it is quantile regression.

In a case-control design, the data at a single variant consist of n_1 cases $\{x_i, \mathbf{z}_i, y_i, d_i = 1\}$, $i = 1, 2, \dots, n_1$, and n_0 controls $\{x_i, \mathbf{z}_i, y_i, d_i = 0\}$, $i = n_1 + 1, n_1 + 2, \dots, n_1 + n_0$. We denote

by $n = n_1 + n_0$ the total sample size. When the secondary phenotype Y is associated with the primary disease D , directly regressing Y against X using case-control samples leads to biased estimation of β_1 . Therefore, we propose a weighted estimating equation-based approach for secondary trait analyses in case-control studies. It utilizes the entire case-control sample and yields consistent estimation of β_1 in the general population.

Weighted estimating equations for the secondary phenotypes in genetic case-control studies

Constructing estimating equations is a common estimation method. The key is to find an estimating function $S(X, Y, \mathbf{Z}, \boldsymbol{\beta})$ such that for randomly selected subjects from the general population, the following equations hold at the true $\boldsymbol{\beta}^*$:

$$E_Y[S(X, Y, \mathbf{Z}, \boldsymbol{\beta}^*)|X, \mathbf{Z}] = 0. \quad (2)$$

In generalized linear models (GLM), the estimating function $S(X, Y, \mathbf{Z}, \boldsymbol{\beta})$ can be constructed as the first derivative of the log-likelihood function, which is known as Fisher's score function. In other regressions that minimize certain loss functions, $S(X, Y, \mathbf{Z}, \boldsymbol{\beta})$ is the first derivative of the corresponding loss function. For example, the estimating function with respect to β_1 is $S(X, Y, \mathbf{Z}, \beta_1) = (Y - \beta_0 - X\beta_1 - \mathbf{Z}^T\boldsymbol{\beta}_2)X$ in linear regression, $S(X, Y, \mathbf{Z}, \beta_1) = (Y - \exp(\beta_0 + X\beta_1 + \mathbf{Z}^T\boldsymbol{\beta}_2)) / (1 + \exp(\beta_0 + X\beta_1 + \mathbf{Z}^T\boldsymbol{\beta}_2))X$ in logistic regression, and $S(X, Y, \mathbf{Z}, \beta_1) = [\tau - I\{Y \leq \beta_0 + X\beta_1 + \mathbf{Z}^T\boldsymbol{\beta}_2\}]X$ at any quantile $\tau \in (0, 1)$ in quantile regression. As we do not have a representative sample of the population, solving Equation 2 directly in a case-control sample is biased. However, conditioning on the disease status D , we can expand the above equation as follows:

$$\begin{aligned} E_Y[S(X, Y, \mathbf{Z}, \boldsymbol{\beta}^*)|X, \mathbf{Z}] \\ = E_Y[S(X, Y, \mathbf{Z}, \boldsymbol{\beta}^*)|X, \mathbf{Z}, D = 0]P(D = 0|X, \mathbf{Z}) \\ + E_Y[S(X, Y, \mathbf{Z}, \boldsymbol{\beta}^*)|X, \mathbf{Z}, D = 1]P(D = 1|X, \mathbf{Z}) = 0. \end{aligned} \quad (3)$$

This expansion provides the basis of constructing the proposed weighted estimating equations. Suppose that for each y_i in the sample, we are able to observe its counterfactual secondary outcomes $\tilde{y}_i$ under the alternative disease status. Specifically, $\tilde{y}_i$ would be the phenotype of the i th case if it is in fact a control, and $\tilde{y}_i$ would be the phenotype of the i th control if it is actually a case. If we are able to observe both y_i and $\tilde{y}_i$'s, we can then construct the unbiased estimating equations following the expanded estimating Equation 3. The sample estimation equations can be written as

$$\begin{aligned} S_n(\boldsymbol{\beta}) = \sum_{i=1}^n [S(x_i, y_i, \mathbf{z}_i, \boldsymbol{\beta})p(d_i|x_i, \mathbf{z}_i) \\ + S(x_i, \tilde{y}_i, \mathbf{z}_i, \boldsymbol{\beta})p(1 - d_i|x_i, \mathbf{z}_i)] = 0, \end{aligned} \quad (4)$$

where the weight $p(d_i|x_i, \mathbf{z}_i)$ is the probability of being the observed disease status given $(x_i, \mathbf{z}_i)$, and $p(1 - d_i|x_i, \mathbf{z}_i)$ is the probability of being the counterfactual disease status. The optimization of estimating Equation 4 can be viewed as a

weighted regression, where weights are $p(d_i|x_i, \mathbf{z}_i)$ for the actual outcomes y_i and $p(1 - d_i|x_i, \mathbf{z}_i)$ for the counterfactual outcomes $\tilde{y}_i$. For this reason, we name our proposed approach the WEE.

One can show that for each summand of Equation 4, its conditional expectation given $(x_i, \mathbf{z}_i, d_i)$ is zero at the true $\boldsymbol{\beta}^*$ and thus constitutes an unbiased estimating equation. Following classical theories for M and Z estimations (theorems 5.7 and 5.9 in Van der Vaart 2000), solving Equation 4, $S_n(\boldsymbol{\beta}) = 0$, leads to the consistent estimation of $\boldsymbol{\beta}$ as $S_n(\boldsymbol{\beta})$ is a consistent estimation function as long as y_i is a random sample conditioning on $(d_i, x_i, \mathbf{z}_i)$. In IPW, the coefficients can be unbiasedly estimated only when y_i given d_i is a random sample. Therefore, the proposed approach is less sensitive to the sampling scheme than the IPW approach. Although the estimating equations involve $P(D|X, \mathbf{Z})$, we are not assuming the disease probability relates only to $(X, \mathbf{Z})$. In reality, the disease risk can relate to Y or other auxiliary variables $\mathbf{W}$ as well, and $p(D|X, \mathbf{Z})$ in Equation 4 can be viewed as the marginal probability given $(X, \mathbf{Z})$; i.e., $p(D|X, \mathbf{Z}) = \int_{y, \mathbf{w}} p(D|X, \mathbf{Z}, y, \mathbf{w}) dF_{(y, \mathbf{w})}(y, \mathbf{w})$, where $F_{(y, \mathbf{w})}$ is the joint distribution of $(y, \mathbf{w})$.

Of course in practice we are unable to observe the counterfactual secondary outcomes. To get around this difficulty, we propose two approaches. The first one is to estimate the expectation of counterfactual estimating functions. When $S(x_i, y_i, \mathbf{z}_i, \boldsymbol{\beta})$ is linear in y_i , we recommend to replace $S(x_i, \tilde{y}_i, \mathbf{z}_i, \boldsymbol{\beta})$ by its conditional expectation over $\tilde{y}_i$. In general for nonlinear estimation functions, we propose to generate pseudo observations from two sets of stratified models. In the next two subsections, we elaborate on the two approaches under the assumption that the probability $p(d_i|x_i, \mathbf{z}_i)$ is known, and an algorithm to estimate $p(d_i|x_i, \mathbf{z}_i)$ is followed.

Estimation approach A: estimating the expected counterfactual estimating function

One can easily show that the following estimating equations, which substitute $S(x_i, \tilde{y}_i, \mathbf{z}_i, \boldsymbol{\beta})$ by its conditional expectation, are unbiased as well:

$$\begin{aligned} S_n(\boldsymbol{\beta}) = \sum_{i=1}^n [S(x_i, y_i, \mathbf{z}_i, \boldsymbol{\beta})p(d_i|x_i, \mathbf{z}_i) \\ + E_{\tilde{y}_i}[S(x_i, \tilde{y}_i, \mathbf{z}_i, \boldsymbol{\beta})|x_i, \mathbf{z}_i]p(1 - d_i|x_i, \mathbf{z}_i)] = 0. \end{aligned} \quad (5)$$

When the estimating function $S(x_i, \tilde{y}_i, \mathbf{z}_i, \boldsymbol{\beta})$ is linear in $\tilde{y}_i$, this approach is particularly appealing since one can simply replace $\tilde{y}_i$ by its conditional mean. In this case, the estimation Equation 5 is equivalent to

$$\begin{aligned} S_n(\boldsymbol{\beta}) = \sum_{i=1}^n [S(x_i, y_i, \mathbf{z}_i, \boldsymbol{\beta})p(d_i|x_i, \mathbf{z}_i) \\ + S\{x_i, E(\tilde{y}_i|x_i, \mathbf{z}_i), \mathbf{z}_i, \boldsymbol{\beta}\}p(1 - d_i|x_i, \mathbf{z}_i)] = 0. \end{aligned} \quad (6)$$

The conditional means $E(\tilde{y}_i|x_i, z_i)$ can be easily estimated from stratified linear regression. Specifically, one can regress y_i against x_i and z_i separately among cases and controls and estimate $E(\tilde{y}_i|x_i, z_i)$ by the predicted value under the alternative disease status model. This way, the estimate can be obtained using one-step optimization, by solving the following equation,

$$\hat{S}_n(\boldsymbol{\beta}) = \sum_{i=1}^n [S(x_i, y_i, z_i, \boldsymbol{\beta})p(d_i|x_i, z_i) + S\{x_i, \hat{E}(\tilde{y}_i|x_i, z_i), z_i, \boldsymbol{\beta}\}p(1-d_i|x_i, z_i)] = 0, \quad (7)$$

where $\hat{E}(\tilde{y}_i|x_i, z_i)$ is the predicted outcome given x_i and z_i under the alternative disease status.

Estimating Equation 5 remains unbiased when $S(x_i, \tilde{y}_i, z_i, \boldsymbol{\beta})$ is nonlinear. However, estimating $E_{\tilde{y}_i}[S(x_i, \tilde{y}_i, z_i, \boldsymbol{\beta})|x_i, z_i]$ in such a case could be computationally undesirable, especially when the model is high-dimensional. We hence propose an alternative approach that is flexible and computationally efficient.

Estimation approach B: generating pseudo counterfactual observations

Under model (1), the linear association between Y and (X, Z) holds among both cases and controls. The regression coefficients, however, can vary between them. Therefore, we propose to fit model (1) separately for cases and controls and use the resulting stratified models to generate pseudo counterfactual observations. Here, we consider two scenarios to illustrate this idea. One is the GLM and the other is quantile function.

Simulating counterfactual outcomes in the GLM

Suppose

$$\hat{g}_d(Y) = \hat{\beta}_{d0} + X\hat{\beta}_{d1} + Z^T\hat{\boldsymbol{\beta}}_{d2}$$

is the stratified estimated models for cases and controls, where $d = 1$ for cases and $d = 0$ for controls. For each observation i , we generate its counterfactual outcome from the estimated model of the alternative disease status g_{1-d} . Specifically, if g is a logit link for a binary secondary trait, then $\hat{\tilde{y}}_i$ is a random draw from a Bernoulli distribution with success probability $\exp\{\hat{g}_{1-d}(y_i)\}/[1 + \exp\{\hat{g}_{1-d}(y_i)\}]$. If g is a log link for a counted secondary trait, then $\hat{\tilde{y}}_i$ is a random draw from a Poisson distribution with $\lambda_i = \exp\{\hat{g}_{1-d}(y_i)\}$.

Simulating counterfactual outcomes in quantile regression

Since quantile regression does not have full parametric likelihood functions, one needs to consider the main model (1) across the entire distribution of Y to simulate counterfactual phenotypes. This joint modeling approach for quantile regression has been described in Wei *et al.* (2006), and the quantile-based counterfactual outcome generations have been described in detail in Wei *et al.* (2015). Here, we

summarize how we have been able to generate counterfactual outcomes based on a grid of evenly spaced quantiles. We let $0 < \tau_1 < \tau_2 < \dots < \tau_k < 1$ denote a set of k evenly spaced quantile levels and $\boldsymbol{\beta}^*(\tau|d) = \{\beta_0^*(\tau|d), \beta_1^*(\tau|d), \beta_2^*(\tau|d)\}$ denote the quantile coefficient functions given disease status $D = d$ such that

$$\boldsymbol{\beta}^*(\tau|d) = \arg \min_{\boldsymbol{\beta}} E_Y[S_{\tau}(X, Y, Z, \boldsymbol{\beta})|X, Z, D = d], \quad (8)$$

for any $\tau \in (0, 1)$. Then $\hat{\tilde{y}}_i$ is simulated following the model-estimated conditional distribution of y_i given (x_i, z_i) and d_i as follows:

1. We estimate the quantile coefficients for $\boldsymbol{\beta}(\tau_k|d)$ in Equation 8 within cases and controls, respectively.
2. To approximate the coefficient process $\boldsymbol{\beta}^*(\tau|d)$, we define $\hat{\boldsymbol{\beta}}(\tau|d)$ to be a piecewise linear function on $[0, 1]$ that concatenates the estimates $\hat{\boldsymbol{\beta}}(\tau_k|d)$ for $0 < \tau_1 < \tau_2 < \dots < \tau_k < 1$ and is subject to the constraint of $\hat{\boldsymbol{\beta}}'(0|d) = \hat{\boldsymbol{\beta}}'(1|d) = 0$.
3. We randomly draw the quantile level u_i from Uniform $(0, 1)$ distribution and simulate the pseudo outcome $\tilde{y}_i$ by $\hat{\tilde{y}}_i = \hat{\beta}_0(u_i|1-d_i) + \hat{\beta}_1(u_i|1-d_i)X + Z^T\hat{\boldsymbol{\beta}}_2(u_i|1-d_i)$.

Stabilizing the coefficients

With $\hat{\tilde{y}}_i$, we construct the sampling estimating equations as

$$\sum_{i=1}^n [S(x_i, y_i, z_i, \boldsymbol{\beta})p(d_i|x_i, z_i) + S(x_i, \hat{\tilde{y}}_i, z_i, \boldsymbol{\beta})p(1-d_i|x_i, z_i)] = 0. \quad (9)$$

As simulating the pseudo counterfactual outcome might introduce variation for small samples, we propose to repeat the procedure above T times to obtain stable estimates. The final estimate is the average of the T estimates; i.e.,

$$\hat{\boldsymbol{\beta}} = T^{-1} \sum_{t=1}^T \hat{\boldsymbol{\beta}}^{(t)}, \quad (10)$$

where $\hat{\boldsymbol{\beta}}^{(t)}$ is the estimated coefficient from the t th set of pseudo outcomes. In our numerical studies, we used T from 1 to 100 and found that the variance of the estimates stabilizes fairly quickly after $T = 10$.

Estimation of $p(d_i|x_i, z_i)$

For the two estimation algorithms described above, we assumed that the conditional disease probability $p(d_i|x_i, z_i)$ was known. In practice, it needs to be estimated. To estimate $p(d_i|x_i, z_i)$, we could use the models from primary analysis or simply assume a logistic model

$$P(D = 1|X, Z) = \frac{\exp(\gamma_0 + X\gamma_1 + Z^T\boldsymbol{\gamma}_2)}{1 + \exp(\gamma_0 + X\gamma_1 + Z^T\boldsymbol{\gamma}_2)}. \quad (11)$$

We can achieve a consistent estimation of the slope parameters γ_1 and $\boldsymbol{\gamma}_2$ by conducting logistic regression in the

case-control sample. The intercept γ_0 needs to be calibrated to match the overall disease prevalence in the general population. Let P_0 denote the known disease prevalence; then we can estimate γ_0 by solving the following equation,

$$P_0 = \int_{X,Z} \frac{\exp(\gamma_0 + X\hat{\gamma}_1 + \mathbf{Z}^T\hat{\gamma}_2)}{1 + \exp(\gamma_0 + X\hat{\gamma}_1 + \mathbf{Z}^T\hat{\gamma}_2)} dF_{XZ}, \quad (12)$$

where F_{XZ} is the joint distribution of X and $\mathbf{Z}$, and $\hat{\gamma}_1$ and $\hat{\gamma}_2$ are the estimated γ_1 and γ_2 from logistic regression. The joint distribution F_{XZ} can be estimated using population databases. If difficult to obtain, we propose a sample version to approximate γ_0 as follows:

$$\hat{\gamma}_0 = \arg \min_{\gamma_0} \left[P_0 - n^{-1} \sum_{i=1}^n \frac{\exp(\gamma_0 + x_i\hat{\gamma}_1 + \mathbf{z}_i^T\hat{\gamma}_2)}{1 + \exp(\gamma_0 + x_i\hat{\gamma}_1 + \mathbf{z}_i^T\hat{\gamma}_2)} \right]^2. \quad (13)$$

When the disease prevalence P_0 or the disease model is misspecified, the resulting estimated coefficients could be slightly biased. The estimation of the coefficients when the $\hat{P}(D|X, \mathbf{Z})$ is misspecified is considered in simulations.

Bootstrap procedure for the confidence intervals and hypothesis tests

In previous sections, we have outlined two estimation algorithms to estimate the parameters in model (1). Although the estimates can be viewed as some form of weighted regressions, the direct output of the Wald test statistics does not take into account the uncertainty from the estimated $\hat{p}(d|x, \mathbf{z})$ and simulated $\hat{y}_i$. Therefore, we propose to use a bootstrap method to obtain the variance-covariance matrices of our proposed estimates. With the bootstrap standard errors, we are able to construct bootstrap confidence intervals and apply Wald test statistics to test the null hypothesis $H_0: \beta_1 = 0$, i.e., whether the genetic variant(s) is (are) associated with the secondary phenotype Y in the general population. The bootstrap procedure is as follows:

1. Bootstrap cases and controls separately to assemble a bootstrap case-control sample. Namely we randomly select n_1 cases from the case sample and n_0 controls from the control sample with replacement.
2. For each bootstrap sample, we reapply the proposed algorithm to obtain bootstrap estimates. For approach A, it includes re-estimating $E_{\tilde{y}_i}[S(x_i, \tilde{y}_i, \mathbf{z}_i, \boldsymbol{\beta})]$ and $p(d|x_i, \mathbf{z}_i)$. For approach B, it includes regenerating pseudo outcomes $\hat{\tilde{y}}_i$ and re-estimating $p(d|x_i, \mathbf{z}_i)$.
3. We repeat steps 1 and 2 B times and then calculate the bootstrap standard error. We use the bootstrap standard error to construct confidence intervals and bootstrap chi-square test statistics for inference.

We evaluated the type I error and power of this bootstrap procedure in simulations and the bootstrap-based inferences are applied to the real data examples.

Simulation Results

Finite sample performance

We use simulations to evaluate the performance of the proposed WEE approach and compare its findings with those of several commonly used methods. We consider the scenario of a preselected SNP with binary and continuous secondary phenotypes in the framework of a case-control study. As quantile regression has been extensively explored in Wei *et al.* (2015), here we focus on its performance in parametric regressions (logistic and linear regressions).

Model settings: As before, we denote that $D = \{0, 1\}$ is the primary disease status, $X = \{0, 1, 2\}$ is the genotype information at the SNP (under an additive model) with minor allele frequency (MAF) = 0.3, and Z is a covariate of interest following a standard normal distribution. The correlation coefficient between X and Z is 0.3. We consider both binary and continuous secondary phenotypes Y . For binary Y , we assume a logistic model as follows:

$$P(Y = 1|X, Z) = \frac{\exp(-1 + 0.2X + 0.1Z)}{1 + \exp(-1 + 0.2X + 0.1Z)}.$$

For continuous Y , we assume the linear model

$$Y = 1 + 0.2X + 0.1Z + \varepsilon,$$

where the error term ε follows $N(0, 1)$. Under the binary secondary phenotype model, the prevalence of Y is $\sim 30\%$. In both models, $\beta_1^* = 0.2$ is the true coefficient of X in predicting Y and is the coefficient of primary interest.

To model the disease probability $P(D|X, Y, Z)$, we consider three possible settings. Setting 1 (logistic setting) assumes the probability of disease follows a logistic model with main effects of X and Y . Similar settings were considered in Lin and Zeng (2009) and Wang and Shete (2011). Setting 2 (interaction setting) extends the logistic setting by including the $X - Y$ interaction. Similar settings were considered in Li and Gail (2012) and Wang and Shete (2012). Finally, setting 3 (piecewise setting) assumes that $P(D|X, Y, Z)$ follows a piecewise linear model instead of logistic regression. The detailed mathematical forms of the disease models are given below.

Setting 1 (logistic setting):

$$\begin{aligned} P(D = 1|X, Y, Z) &= \frac{\exp(\gamma_0 + 0.3X + \log(2)Y + \log(2)Z)}{1 + \exp(\gamma_0 + 0.3X + \log(2)Y + \log(2)Z)}. \end{aligned}$$

Setting 2 (interaction setting):

$$\begin{aligned} P(D = 1|X, Y, Z) &= \frac{\exp(\gamma_0 + 0.3X + \log(2)Y + \log(2)Z + 0.2XY)}{1 + \exp(\gamma_0 + 0.3X + \log(2)Y + \log(2)Z + 0.2XY)}. \end{aligned}$$

Setting 3 (piecewise setting):

$$Q = 0.3X + \log(2)Y + \log(2)Z$$

$$P(D = 1|X, Y, Z) = \begin{cases} \gamma_0, & Q \leq U_1 \\ \gamma_0 + 0.1 \frac{Q - U_1}{U_2 - U_1}, & U_1 < Q \leq U_2 \\ \gamma_0 + 0.1, & Q > U_2, \end{cases}$$

where U_1 and U_2 are the 0.25th and 0.75th quantiles of the $0.3X + \log(2)Y + \log(2)Z$, respectively.

In all the disease models above, we assume that the prevalence of the primary disease (P_0) is 10%. The intercepts γ_0 in these models are selected to match the overall disease prevalence in the population. For each of the model settings above, we generate a large number of observations ($N = 500,000$), which we treat as a general population.

Sampling schemes: We consider three sampling schemes to select the case-control samples. First, we consider a simple sampling scheme in which cases and controls are randomly drawn into the sample, creating the selection probability depends only on the disease status. This is the simplest available sampling design, but the collection of data is often difficult, especially when the sample size is large. Second, we consider the convenient sampling schemes to collect data that researchers often encounter in case-control studies. For example, in a stroke GWAS (Cornelis *et al.* 2010), cases were selected from imaging-confirmed cases from seven clinical centers, while controls were selected using existing healthy subjects in an acute myocardial infarction study with similar recruitment criteria. Data under this sampling scheme are easy to collect, but the confounding factors associated with selection probabilities may not be fully captured. We then evaluate the performance of the WEE and existing methods in the presence of unmeasured confounders. Finally, we consider the stratified sampling scheme, which is often used in large-scale case-control studies to under- or oversample subjects with certain characteristics. This sampling scheme can largely improve the estimation efficiency for the small subgroups. For example, the National Maternal and Infant Health Survey oversampled the infants born with low birthweight (≤ 2500 g) and very low birthweight (≤ 1500 g) to study the long-term and short-term health outcomes of these infants. This is a common scenario when IPW fails because certain subjects carrying large weights may dominate the estimates. Specifically, we select 2000 cases and 2000 controls into the sample in one of the following ways:

Setting 1 (random sample): Subjects are randomly selected into the sample.

Setting 2 (convenience sample): Assume V is an uncaptured variable associated with (X, Y, Z) such that

$$P(V = 1|X, Y, Z) = \frac{\exp(\gamma_0 + 0.3X + \log(2)Y + \log(2)Z)}{1 + \exp(\gamma_0 + 0.3X + \log(2)Y + \log(2)Z)},$$

and only the subjects with $V = 1$ are selected into the sample. Setting 3 (stratified sample): $V \sim N(0, 1)$ and $V \perp (X, Y, Z, D)$.

Subjects with $V > 0$ are nine times more likely to be selected into the sample than subjects with $V < 0$.

Comparison methods: We estimated the coefficients under the random sample of different models above, using the following methods: (1) regression using cases only, (2) regression using controls only, (3) regression using a combined case-control sample, (4) regression using a disease status stratified case-control sample, (5) IPW, (6) SPML, and (7) the proposed WEE approach. In the IPW approach, we model cases and controls separately and weight them inversely to their probability of selection $w_d = P(D = d)/n_d$ for cases ($d = 1$) and controls ($d = 0$). The estimating equation of IPW is as follows: $\sum_{d=0}^1 w_d \sum_{i:D_i=d} S(X_i, Y_i, Z_i, \beta) = 0$. For the SPML approach, we use the external SPREG software provided by the authors (<http://dlin.web.unc.edu/software/spreg-2/>). In the WEE approach, we use estimating approach A of taking the conditional expectation of the counterfactual score function in the linear regression, which solves through a simple one-step optimization; we use estimating approach B of generating pseudo observations in logistic regression. We vary the T values from 1 to 100 (T is the number of pseudosamples generated) and compare their estimates. We assume we have no prior knowledge of F_{XZ} and the weight $P(D|X, Z)$ is estimated through the sample version equation (Equation 13). We then further compare the WEE with the IPW and SPML approaches in both the convenient and stratified samples. As outlined in the Introduction, IPW is the most simple and robust method and SPML is the most efficient and popular method in the current available secondary analysis literature.

Results and discussions: Table 1 summarizes the relative bias, standard error, and mean square error of the estimated coefficients β_1 based on 500 Monte Carlo replicates from random samples using all methods. According to Table 1, the classical methods, including regressions applied to the cases only, controls only, combined case-control sample directly, and combined case-control sample adjusting for primary disease status, are all biased. Hence, without appropriate adjustment, classical methods provide biased estimation for the $X - Y$ association in genetic case-control data.

The proposed WEE approach performs well in correcting such bias in all the settings we considered. In logistic regression, we generated pseudo counterfactual observations. The estimated coefficients are unbiased even with $T = 1$. The standard errors do decrease slightly as T increases, but they quickly stabilize after $T = 10$. Therefore, we conclude that a relatively small number of imputations is enough to reach the

Table 1 The relative bias (RB), standard error (SE), and mean square error (MSE) of the estimated coefficient β_1 under a *simple sampling scheme*

Model	Method	Logistic			Interaction			Piecewise		
		RB (%)	SE	MSE $\times n$	RB (%)	SE	MSE $\times n$	RB (%)	SE	MSE $\times n$
LogR	Case	−13.3	0.067	10.2	66.3	0.064	42.4	−36.5	0.072	21.4
	Control	−10.2	0.085	15.3	−30.1	0.081	20.9	3.5	0.081	13.3
	CC	10.4	0.050	5.9	61.4	0.050	34.4	−12.0	0.056	7.5
	Adj CC	−12.1	0.051	6.2	26.9	0.051	10.6	−18.1	0.056	9.1
	IPW	0.3	0.072	10.5	2.0	0.069	9.5	1.1	0.073	10.5
	SPML	0.5	0.050	5.0	46.5	0.050	21.7	−15.5	0.056	8.3
	WEE ($T = 1$)	−0.4	0.082	13.6	−0.8	0.076	11.7	0.9	0.080	12.9
	WEE ($T = 10$)	−0.5	0.074	10.8	−0.5	0.071	10.0	0.9	0.073	10.8
	WEE ($T = 100$)	−0.5	0.073	10.6	−0.4	0.070	9.8	0.6	0.073	10.6
LR	Case	−24.8	0.031	6.8	13.3	0.032	3.6	−27.6	0.032	7.9
	Control	−10.6	0.037	3.5	−34.7	0.035	11.8	0.5	0.036	2.6
	CC	12.8	0.024	2.5	52.7	0.025	23.8	−8.6	0.024	1.7
	Adj CC	−18.6	0.023	3.7	−8.4	0.024	1.6	−14.2	0.024	2.6
	IPW	1.6	0.032	2.1	−0.8	0.031	1.9	−0.5	0.032	2.0
	SPML	0.2	0.026	1.3	28.3	0.024	7.7	−12.1	0.024	2.2
	WEE	0.6	0.033	2.1	0.6	0.033	2.1	−0.7	0.032	2.1

The true value $\beta_1^* = 0.2$. LogR, logistic regression; LR, linear regression; Case, unadjusted regression using case sample only; Control, unadjusted regression using control sample only; CC, unadjusted regression using both case and control samples; Adj CC, regression using both case and control samples, adjusting for primary disease status; IPW, inverse probability weighting regression; SPML, semiparametric maximum-likelihood regression; WEE (T), the proposed weighted estimating equation approach estimated by generating pseudo counterfactual observations with T replicates; WEE, the proposed weighted estimating equation approach estimated by solving its conditional expectation.

optimal efficiency of this approach, and therefore it is computationally efficient. In addition, we assumed the working model as $P(D|X, Z) = \exp(\gamma_0 + \gamma_1 X + \gamma_2 Z) / (1 + \exp(\gamma_0 + \gamma_1 X + \gamma_2 Z))$ in estimation, which is not affected by the $Y - D$ associations. It is different from the one used to generate the data; the generating model is based on three $P(D|X, Y, Z)$ settings, including logistic, interaction, and piecewise settings. Even under the misspecified $P(D|X, Z)$, the proposed estimating equation-based approach performs fairly well, indicating that it is quite robust to the $P(D|X, Z)$ model misspecification. Additional scenarios to test the robustness boundary of $\hat{P}(D|X, Z)$ are described in later sections.

The IPW also produces fairly accurate estimates in all models and demonstrates comparable efficiency in comparison with the WEE in Table 1. The calculation of the IPW method, however, requires information on the sampling scheme. Under random samples, where the selection of cases and controls depends solely on the disease status, it is similar to using the disease prevalence to calculate $p(d_i|x_i, z_i)$ in the WEE. Therefore, it is not surprising to observe comparable performance in this scenario. Table 2 further compares IPW with the WEE under complex sampling schemes and shows that IPW is less robust and efficient than the WEE in these scenarios. Under convenient sampling schemes with unadjusted confounding factors, every method contains some biases. The biases are controlled within 7% in the WEE, but they can reach up to 25% in IPW. This is because the WEE is less affected by the confounding factors as long as Y given (X, Z) is a representative sample of the population. Under the stratified sampling scheme, the IPW estimates suffer from inflated variance, because subjects with large sampling weights dominate the estimates. However, subjects with

similar (X, Z) values contribute similar weights in the WEE, which therefore is robust to the stratification. We further compare their power under this scenario in the next section.

The SPML approach provides unbiased and relatively efficient estimations when the linear logistic model assumption is satisfied but introduces biases when this assumption is violated. Specifically, the SPML estimate is the most efficient of all methods we considered for random samples under the logistic setting. Under interaction and piecewise settings, where the linear logistic model assumption is violated, the SPML estimates are very sensitive to the model misspecification and contain considerable bias. As SPML is more sensitive to the underlying disease model than the WEE is, its performance under convenient samples may be worse than that of the WEE even for the linear logistic disease model. Its performance under stratified samples is comparable to that under random samples, as, similar to the WEE, SPML does not use sampling weights that might dominate the estimates.

Type I error estimates and power comparisons

In this section, we further investigate the type I error and power of the WEE approach with the bootstrap procedure and compare it with IPW and SPML for the primary hypothesis $H_0: \beta_1 = 0$. For type I error comparisons, we consider the association between a binary Y and a single preselected SNP with no covariates under random samples. The MAF of the SNP ranges from 10% to 50%, and the primary disease models follow logistic, interaction, or piecewise settings. We simulate 100,000 Monte Carlo samples with 2000 cases and 2000 controls. Table 3 summarizes the type I errors of IPW, SPML, and WEE methods at the α levels of 0.05, 0.01, and 0.001. According to Table 3, both the WEE method and IPW

Table 2 The relative bias (RB), standard error (SE), and mean square error (MSE) of the estimated coefficient β_1 under *complex sampling schemes*

Sample (model)	Method	Logistic			Interaction			Piecewise		
		RB (%)	SE	MSE $\times n$	RB (%)	SE	MSE $\times n$	RB (%)	SE	MSE $\times n$
Convenient (LogR)	IPW	−24.3	0.081	17.7	−21.5	0.077	15.6	−24.2	0.077	16.6
	SPML	−4.7	0.053	5.7	14.3	0.057	8.1	−48.4	0.055	24.7
	WEE ($T = 10$)	−6.9	0.047	4.8	−6.5	0.045	4.4	−4.7	0.045	4.2
Convenient (LR)	IPW	9.9	0.032	2.9	11.3	0.032	3.1	6.1	0.033	2.4
	SPML	7.6	0.033	3.6	−20.3	0.041	6.7	−40.6	0.023	14.2
	WEE	5.8	0.033	2.4	0.9	0.032	2.1	4.8	0.033	2.3
Stratified (LogR)	IPW	4.7	0.119	28.4	4.5	0.114	26.0	−1.5	0.112	25.0
	SPML	3.6	0.053	5.6	39.8	0.052	17.3	−12.9	0.054	7.1
	WEE ($T = 10$)	0.4	0.073	10.7	−1.4	0.072	10.5	−1.9	0.072	10.4
Stratified (LR)	IPW	0.5	0.053	5.7	−0.7	0.048	4.6	0.0	0.052	5.5
	SPML	3.9	0.025	1.4	27.4	0.050	5.1	−12.2	0.024	2.3
	WEE	1.2	0.033	2.1	0.1	0.031	2.1	−0.7	0.031	1.9

The true value $\beta_1^* = 0.2$. LogR, logistic regression; LR, linear regression; IPW, inverse probability weighting regression; WEE (T), the proposed weighted estimating equation approach estimated by generating pseudo counterfactual observations with T replicates; WEE, the proposed weighted estimating equation approach estimated by solving its conditional expectation.

produce the correct type I errors in all the settings, while the SPML method results in largely inflated type I error in the interaction and the piecewise settings due to its deviation from the linear logistic model assumption.

For the scenarios in which the novel methods hold valid type I errors, we compare their power under two values for the narrow heritability $h^2 = 0.01$ and 0.02 , and the $X - Y$ association coefficient $\beta = \sqrt{h^2/2\text{MAF}(1 - \text{MAF})}$ ranges between 0.14 and 0.33 as a result. The power at significance level $\alpha = 0.05$ for 1000 Monte Carlo samples with 2000 cases and 2000 controls using the random and stratified sampling schemes is presented in Table 4. Under random samples, the WEE and IPW demonstrate similar power regardless of the underlying disease models. The SPML is also more powerful for detecting the $X - Y$ associations when the underlying disease model $P(D|X, Y)$ is linear logistic, but its type I error blows up when the underlying models do not satisfy its linear logistic assumption. Under the stratified sampling scheme, where subjects with large sampling weights may dominate the estimates in IPW, the WEE proves far more powerful than IPW.

The performance under biased estimated $\hat{P}(D|X, Z)$

The proposed estimates made two assumptions in the estimation of $P(D|X, Z)$ that the primary disease prevalence is known as P_0 and the estimated $P(D|X, Z)$ follows a linear logistic model as follows:

$$P(D = 1|X, Z) = \frac{\exp(\gamma_0 + X\gamma_1 + Z^T\gamma_2)}{1 + \exp(\gamma_0 + X\gamma_1 + Z^T\gamma_2)}. \quad (14)$$

The proposed estimates were obtained in previous simulations under model misspecification (the true generating models were logistic, interaction, and piecewise settings), and the performances indicate that the WEE is fairly robust with linear logistic assumption. In this section, we further investigate the robustness of our proposed method when the P_0 is incorrectly

estimated. The disease prevalence is often estimated from cohort studies or literature and can also be biased. Instead of using the true disease prevalence-generated $P_0 = 10\%$, we assume that we obtain the misspecified $\hat{P}_0$ value ranging from 5% to 20% for estimation under the logistic setting. Table 5 shows the relative bias, standard error, and mean square error of the estimated coefficients from 500 Monte Carlo replicates with various $\hat{P}_0$ values used for estimation. We find that while the estimation bias does increase slowly when the deviation from the true prevalence increases, even doubling the disease prevalence incurs biases of $<7\%$. We conclude that the resulting estimates are fairly robust against the misspecified prevalence.

Real Data Analysis

In this section, we apply the proposed WEE approach to an asthma case-control GWAS from the New York University Bellevue Asthma Registry (NYUBAR) (Liu *et al.* 2011). The study consisted of 387 asthmatics and 212 healthy controls, genotyped 10 tag SNPs at the thymic stromal lymphopoietin (TSLP) gene, and identified the association between the TSLP gene and asthma in the primary analysis. To do so, the study controlled for a number of demographic and clinical variables such as age, gender, race, smoking status, BMI, forced expiratory volume in 1 second (FEV₁), and IgE level. As an ongoing non-National Institutes of Health funded study, the asthma cohort data are not currently available to the public.

To illustrate the proposed approach, we consider two secondary phenotypes: one is a binary secondary phenotype of overweight status, and the other is a continuous secondary phenotype of serum IgE levels. Overweight is defined as body mass index >25 . A meta-analysis by Flaherman and Rutherford (2006) combining 402 studies from 1966 to October 2004 found that the effect of high body weight during middle childhood showed a 50% increase in relative risk (relative

Table 3 Type I error of WEE approach in comparison with IPW and SPML for a preselected SNP

Method	MAF	Logistic			Interaction			Piecewise		
		0.05	0.01	0.001	0.05	0.01	0.001	0.05	0.01	0.001
IPW	0.1	0.05091	0.01078	0.00145	0.05066	0.01080	0.00136	0.05064	0.01077	0.00113
	0.2	0.05169	0.01116	0.00131	0.05214	0.01123	0.00145	0.05285	0.01130	0.00124
	0.3	0.05301	0.01166	0.00142	0.05142	0.01045	0.00102	0.05233	0.01124	0.00122
	0.4	0.05113	0.01106	0.00126	0.05316	0.01129	0.00138	0.05378	0.01187	0.00156
	0.5	0.05233	0.01120	0.00127	0.05317	0.01117	0.00125	0.05292	0.01159	0.00144
SPML	0.1	0.05050	0.00930	0.00090	0.25043	0.10064	0.02311	0.90880	0.75590	0.46780
	0.2	0.05710	0.01190	0.00140	0.37414	0.17526	0.04829	0.97870	0.91760	0.74050
	0.3	0.05900	0.01430	0.00150	0.42709	0.21057	0.06665	0.98750	0.94040	0.79260
	0.4	0.06660	0.01340	0.00180	0.44216	0.22318	0.06905	0.98290	0.93490	0.78070
	0.5	0.06800	0.01541	0.00201	0.42486	0.21007	0.06234	0.99270	0.96570	0.85900
WEE	0.1	0.05018	0.01083	0.00125	0.05098	0.01072	0.00128	0.05018	0.01083	0.00125
	0.2	0.05260	0.01133	0.00117	0.05247	0.01135	0.00138	0.05260	0.01133	0.00117
	0.3	0.05178	0.01157	0.00135	0.05151	0.01070	0.00108	0.05178	0.01157	0.00135
	0.4	0.05360	0.01176	0.00153	0.05271	0.01115	0.00128	0.05360	0.01176	0.00153
	0.5	0.05336	0.01205	0.00148	0.05228	0.01068	0.00134	0.05336	0.01205	0.00148

IPW, inverse probability weighting logistic regression; SPML, semiparametric maximum-likelihood-based logistic regression; WEE, the proposed weighted estimating equation approach estimated by generating pseudo counterfactual observations with $T = 10$ replicates.

risk = 1.5, 95% C.I. = 1.2–1.8) of having subsequent asthma. The NYUBAR data set is consistent with their findings with an odds ratio (OR) of having asthma for overweight observations of 1.66 (P -value = 0.006; 95% C.I. = 1.15–2.38) compared to the normal counterparts. As a result, the commonly used classical secondary analysis methods may provide biased estimation. We sought to apply more appropriate approaches for estimation.

IgE is a class of antibody that mediates the immune responses in the pathogenesis of allergic asthma (Burrows *et al.* 1989). An allergen-specific IgE level >0.35 kilo-international units (kIU)/liter is considered positive. Elevated IgE is associated with many allergic diseases, such as allergic rhinitis, peanut allergy, latex sensitivity, atopic dermatitis, and chronic urticaria (Morjaria and Polosa 2009). The secondary analysis of IgE leads to better understanding of the mechanism by which TSLP influences risk to asthma and other allergic diseases. As serum IgE level is approximately normally distributed among cases and controls after log transformation, we first apply linear regression for estimation. In addition, as elevated IgE level instead of mean IgE level plays an essential role in allergic diseases, we also consider the quantile regression approach to further investigate the genetic association with upper quantiles of the log serum IgE levels. Given that the log serum IgE level is strongly associated with asthma in the data set (OR = 1.4; P -value = $6.7E-9$; 95% C.I. = 1.26–1.58), the commonly used methods may also provide biased estimates of the association between the TSLP gene and log serum IgE level, and we analyze them with novel approaches.

Logistic regression for overweight status

We denote $D = \{0, 1\}$ as the primary case–control status of asthma, $X = \{0, 1, 2\}$ as the minor allele count for each of the 10 TSLP SNPs, $Y = \{0, 1\}$ as the binary secondary trait of

whether a person is overweight, and Z as a continuous variable of the propensity score (Guo and Fraser 2010) derived from a set of covariates including age, gender, smoking status, FEV₁, and the first principal component score from 213 ancestry informative markers (Pritchard *et al.* 2000). Then the logistic model we consider is as follows:

$$Y = \frac{\exp(\beta_0 + \beta_1 X + \beta_2 Z)}{1 + \exp(\beta_0 + \beta_1 X + \beta_2 Z)}.$$

Three approaches are used to estimate the coefficient β_1 : IPW, SPML, and WEE. For this regression and for the mean and quantile regressions in later sections, we calculate the overall asthma prevalence as 10.1% based on six birth cohort studies, and this information is used to approximate selection probability in IPW and estimate $P(D|X, Y, Z)$ in SPML and $P(D|X, Z)$ in the WEE. The population distribution of F_{XZ} is unknown and the weight $P(D|X, Z)$ is estimated through the sample version equation (Equation 13). The standard errors and P -values were calculated using bootstrap; *i.e.*, we bootstrap cases and controls separately and reapply the entire estimating procedure to the bootstrap case–control sample.

The resulting estimated coefficients of the 10 tag SNPs in the TSLP gene and overweight are summarized in Table 6. Before adjusting for multiple comparisons, both IPW and the WEE are able to identify two SNPs (rs2289276 and rs11466741) at the α -level 0.05, while SPML fails to detect any SNPs. The point estimates of the WEE and IPW are comparable, but the standard errors of the WEE are smaller, resulting in smaller P -values. Specifically, the P -values of SNP rs11466741 are 0.008 and 0.047 in the WEE and IPW, respectively.

SPML generates very different estimates from IPW and the WEE at many SNPs. As we mentioned in the Introduction, IPW is widely known as a robust approach and SPML is efficient but potentially biased with misspecified $P(D|X, Y, Z)$. Therefore, we further tested the $X - Y$ interactions in $P(D|X, Y, Z)$

Table 4 Power of the WEE approach in comparison with IPW and SPML under random samples and stratified samples ($\alpha = 0.05$)

Method	MAF	Basecase		Interaction		Piecewise	
		$h^2 = 0.01$	$h^2 = 0.02$	$h^2 = 0.01$	$h^2 = 0.02$	$h^2 = 0.01$	$h^2 = 0.02$
Random sample							
IPW	0.1	0.666	0.922	0.652	0.914	0.645	0.909
	0.2	0.640	0.901	0.661	0.890	0.679	0.923
	0.3	0.631	0.921	0.617	0.903	0.675	0.937
	0.4	0.604	0.902	0.627	0.915	0.683	0.933
	0.5	0.658	0.919	0.614	0.891	0.688	0.921
SPML	0.1	0.832	0.990	<i>a</i>	<i>a</i>	<i>a</i>	<i>a</i>
	0.2	0.774	0.976	<i>a</i>	<i>a</i>	<i>a</i>	<i>a</i>
	0.3	0.786	0.980	<i>a</i>	<i>a</i>	<i>a</i>	<i>a</i>
	0.4	0.770	0.976	<i>a</i>	<i>a</i>	<i>a</i>	<i>a</i>
	0.5	0.748	0.972	<i>a</i>	<i>a</i>	<i>a</i>	<i>a</i>
WEE	0.1	0.659	0.914	0.647	0.916	0.642	0.892
	0.2	0.620	0.894	0.633	0.891	0.669	0.918
	0.3	0.620	0.912	0.610	0.902	0.663	0.932
	0.4	0.598	0.900	0.604	0.902	0.657	0.926
	0.5	0.650	0.914	0.611	0.893	0.683	0.921
Stratified sample							
IPW	0.1	0.314	0.508	0.303	0.509	0.276	0.493
	0.2	0.291	0.506	0.307	0.487	0.302	0.525
	0.3	0.309	0.540	0.288	0.501	0.312	0.549
	0.4	0.285	0.498	0.274	0.461	0.293	0.550
	0.5	0.299	0.532	0.252	0.507	0.307	0.552
SPML	0.1	0.866	0.992	<i>a</i>	<i>a</i>	<i>a</i>	<i>a</i>
	0.2	0.796	0.984	<i>a</i>	<i>a</i>	<i>a</i>	<i>a</i>
	0.3	0.740	0.982	<i>a</i>	<i>a</i>	<i>a</i>	<i>a</i>
	0.4	0.776	0.986	<i>a</i>	<i>a</i>	<i>a</i>	<i>a</i>
	0.5	0.780	0.970	<i>a</i>	<i>a</i>	<i>a</i>	<i>a</i>
WEE	0.1	0.662	0.920	0.628	0.930	0.642	0.870
	0.2	0.622	0.904	0.660	0.892	0.660	0.918
	0.3	0.618	0.922	0.608	0.888	0.656	0.926
	0.4	0.590	0.908	0.582	0.918	0.642	0.916
	0.5	0.652	0.906	0.596	0.890	0.668	0.918

IPW, inverse probability weighting logistic regression; SPML, semiparametric maximum-likelihood-based logistic regression; WEE, the proposed weighted estimating equation approach estimated by generating pseudo counterfactual observations with $T = 10$ replicates.

^a The power is not given due to the inflated type I error.

to understand the underlying models, and some of the interactive effects are significant (SNPs rs2289278 and rs10035870). Therefore, we believe that the SPML approach contains substantial biases due to the violation of the model assumptions, and the WEE as well as IPW provides relatively unbiased estimations.

In summary, this work demonstrated that the WEE approach combines the advantages of IPW and SPML estimators in that it is robust and fairly efficient in estimating the marker-secondary trait associations.

Mean regression for serum IgE level

In this example, we let Y denote a continuous secondary trait of the log serum IgE level. As in the case before, D is the case-control status of asthma, X is the minor allele count for each of the 10 TSLP SNPs, and Z is the propensity score of the covariates. We then consider the IPW, SPML, and WEE approaches under the following linear regression:

$$Y = \beta_0 + \beta_1 X + \beta_2 Z.$$

The resulting estimated coefficients are summarized in Table 7. The WEE identifies three SNPs (rs11466741, rs11466743, and rs10035870) at an α -level of 0.05 before adjusting for multiple comparison, IPW detects one significant SNP rs10035870, and SPML fails to detect any SNP. As in the logistic regression for the overweight status, the point estimates of the WEE and IPW are comparable, and the WEE is able to detect more SNPs because it is more efficient than IPW. SPML not only provides biased estimates due to the violation of its assumption on $P(D|X, Y, Z)$, but also fails to improve the efficiency in comparison with the WEE. The WEE provides the most robust and efficient estimation of the associations between the TSLP gene and serum IgE level.

Quantile regression for serum IgE level

Elevated serum IgE levels instead of the mean serum IgE levels contribute to the allergic effects. Therefore, we also consider quantile regression for the secondary analysis to identify the TSLP variants that are associated with the upper quantiles of the serum IgE level. In addition, we use quantile regression to

Table 5 The sensitivity of the WEE approach to the misspecifying primary disease prevalence $\hat{P}_0$

$\hat{P}_0$	Logistic model			Linear model		
	RB (%)	SE	MSE $\times n$	RB (%)	SE	MSE $\times n$
$P_0/2$	-3.5	0.075	11.3	-6.0	0.032	2.4
$P_0/1.5$	-2.0	0.073	10.7	-4.3	0.032	2.1
$P_0/1.2$	-0.7	0.071	10.2	-2.7	0.031	2.0
P_0	-0.5	0.074	10.8	-1.1	0.030	1.9
$1.2P_0$	2.1	0.068	9.3	0.6	0.030	1.8
$1.5P_0$	4.0	0.066	8.7	2.9	0.029	1.7
$2P_0$	6.8	0.062	8.0	6.2	0.028	1.8

$P_0 = 10\%$ is the true disease prevalence.

deepen our knowledge from mean regression on how SNPs affect the location, scale, and distribution of serum IgE levels. The quantile model we consider is

$$Q_Y(\tau) = \beta_{0,\tau} + \beta_{1,\tau}X + \beta_{2,\tau}Z,$$

where X is the minor allele count for each of the 10 TSLP SNPs, Z is a continuous variable of propensity score developed from covariates, and Y is the log serum IgE level. To evaluate the effects of the TSLP gene variants on different levels of IgE, we estimate the model at quantile levels of 0.15, 0.25, 0.5, 0.75, and 0.85, respectively. Two approaches are used to estimate the coefficient β_1 , IPW and the WEE, as the SPML approach is likelihood based and cannot be applied to nonparametric regressions.

The resulting estimated quantile coefficients are summarized in Table 8. The estimated quantile coefficients from the two approaches are comparable. However, the bootstrap standard errors of the IPW estimates are much bigger than the ones from the WEE. Consequently, the WEE estimates are more powerful in detecting the quantile associations.

For the three SNPs that are significant in mean regression in Table 7, their results from quantile regressions also indicate significant associations, and these associations remain significant even after a conservative Bonferroni correction for estimating different quantile levels and the number of SNPs. Moreover, quantile regression is able to detect the SNPs that are associated only with the upper quantile but not the mean of serum IgE level. Specifically, the WEE shows that SNPs rs2289276, rs2289278, rs2289277, and rs11466750 have significant associations with the 75th quantile of log serum IgE level; however, the mean regression did not indicate significant associations, illustrating the potential for the new approach to discover new associations.

Quantile analysis is able to present a comprehensive picture on the effects of the SNPs on the entire distribution of serum IgE level. To obtain a more complete picture, we estimate the quantile coefficients on a fine grid of quantile levels. In Figure 1, A and B, we plotted the estimated conditional distribution functions with different genotypes at SNPs rs10035870 and rs11466743, respectively. Specifically, the solid curve in Figure 1A is the estimated quantile function for the individuals with genotype AA at rs10035870, and the dashed line is for the individuals with genotype AG/GG. In

Table 6 Estimated mean allelic effects on overweight status in logistic regression

SNP	Method	Est.	SE	P-value
rs2289276	IPW	-0.75	0.28	0.007
	SPML	-0.24	0.19	0.197
	WEE	-0.83	0.30	0.006
rs1898671	IPW	0.27	0.31	0.371
	SPML	0.17	0.19	0.385
	WEE	0.22	0.28	0.425
rs11466741	IPW	-0.58	0.29	0.047
	SPML	-0.18	0.19	0.322
	WEE	-0.68	0.26	0.008
rs11466743	IPW	-0.51	0.71	0.474
	SPML	-0.20	0.46	0.657
	WEE	-0.57	1.03	0.578
rs2289277	IPW	-0.54	0.33	0.102
	SPML	-0.13	0.20	0.525
	WEE	-0.53	0.34	0.113
rs2289278	IPW	-0.28	0.39	0.462
	SPML	-0.07	0.25	0.771
	WEE	-0.43	0.39	0.266
rs11241090	IPW	0.54	0.54	0.319
	SPML	0.55	0.35	0.117
	WEE	0.66	0.54	0.218
rs10035870	IPW	-0.64	0.63	0.307
	SPML	-0.07	0.36	0.837
	WEE	-0.77	0.54	0.155
rs11466749	IPW	-0.04	0.35	0.903
	SPML	-0.07	0.23	0.775
	WEE	-0.04	0.31	0.886
rs11466750	IPW	0.06	0.32	0.848
	SPML	0.01	0.21	0.981
	WEE	0.13	0.32	0.681

IPW, inverse probability weighting logistic regression; SPML, semiparametric maximum-likelihood logistic regression; WEE, the proposed weighted estimating equation approach estimated by simulating pseudo observations; Est., estimation.

Figure 1B, the solid curve is the estimated quantile function for genotype GG at rs11466743, and the dashed line is for genotype AG/AA. Both SNPs were found to have a significant impact on the distribution of serum IgE level. SNP rs10035870 has a strong positive effect on the entire distribution of serum IgE level, and thus subjects with the major allele at rs10035870 tend to have a higher serum IgE level in general. In contrast, SNP rs11466743 has a strong impact only on the median and upper quantiles, but has little effect on the lower quantiles of serum IgE level. As indicated in Figure 1B, the subjects with genotype AG/AA in rs11466743 are less likely to have a very high serum IgE level compared to those with genotype GG; however, they have an equal chance to have a low serum IgE level.

In summary, the quantile regression approach demonstrates two attractive features in marker-secondary trait analysis. First, it is able to identify additional SNPs that are associated only with certain quantiles of Y . Second, it describes a complete picture of the entire Y distribution where the associations exist. Given these findings it is clear that existing approaches are either inefficient for this type of regression (IPW) or unable to perform (SPML). The proposed WEE approach can supplement the secondary analysis by facilitating this type of regression.

Table 7 Estimated mean allelic effects on log serum IgE level in linear regression

SNP	Method	Est.	SE	P-value
rs2289276	IPW	0.09	0.14	0.500
	SPML	0.06	0.10	0.569
	WEE	0.10	0.08	0.252
rs1898671	IPW	−0.12	0.17	0.505
	SPML	−0.12	0.10	0.262
	WEE	−0.11	0.13	0.380
rs11466741	IPW	0.20	0.12	0.112
	SPML	0.03	0.09	0.721
	WEE	0.20	0.09	0.032
rs11466743	IPW	−0.61	0.34	0.073
	SPML	−0.25	0.29	0.382
	WEE	−0.61	0.29	0.034
rs2289277	IPW	0.13	0.12	0.282
	SPML	0.03	0.09	0.739
	WEE	0.13	0.09	0.165
rs2289278	IPW	−0.29	0.22	0.193
	SPML	0.05	0.16	0.764
	WEE	−0.29	0.16	0.077
rs11241090	IPW	0.14	0.33	0.658
	SPML	0.20	0.22	0.377
	WEE	0.14	0.24	0.554
rs10035870	IPW	0.63	0.28	0.024
	SPML	0.03	0.21	0.904
	WEE	0.63	0.22	0.004
rs11466749	IPW	0.07	0.24	0.779
	SPML	0.14	0.15	0.343
	WEE	0.07	0.16	0.678
rs11466750	IPW	−0.05	0.16	0.748
	SPML	0.08	0.12	0.484
	WEE	−0.05	0.13	0.671

IPW, inverse probability weighting linear regression; SPML, semiparametric maximum-likelihood linear regression; WEE, the proposed weighted estimating equation approach estimated by solving its conditional expectation; Est., estimation.

Discussion

In this article, we propose a general framework of weighted estimating equations that provides unbiased estimation of the genetic associations with secondary phenotypes in case-control designs. It enjoys a number of attractive properties in the following aspects.

First, its framework is flexible to accommodate various types of single secondary phenotypes and regressions. We illustrated the WEE, using logistic regression for binary outcomes and linear regression and quantile regression for continuous outcomes. Moreover, with appropriately selected estimating functions, the WEE can be applied to many other types of outcomes, such as ordinal, nominal, count, and time-to-event traits. An example of the application of the WEE for survival analysis is provided in the *Appendix*.

Second, the WEE can easily accommodate multiple SNPs and covariates at the same time. Although the approach presented for GWAS data focuses primarily on a single SNP and a few common covariates including population substructure, it can be applied to a much wider range of scenarios for the secondary analysis. For example, rare variants can be aggregated in a region, and the WEE can be applied to the aggregated data for the detection of rare variant effects in

sequencing studies. It can also be incorporated into high-dimensional data analysis such as conducting variable selection from a large number of SNPs and covariates, using cross-validation or penalty functions like Lasso.

Third, the WEE does not make any assumptions on the $Y - D$ association, which is more general than most likelihood-based approaches. As shown in the simulation studies, the WEE yields unbiased $X - Y$ association estimation regardless of the $Y - D$ association.

Fourth, the WEE is not sensitive to sampling schemes. Although the IPW approach is a simple and flexible method that works for any model, it requires knowing additional information on sampling probabilities. The resulting estimates may not be robust or efficient under its sampling schemes, such as under the existence of confounding factors and overweighted observations.

Finally, the WEE is computationally simple and straightforward. The essence of the new estimating equation is weighted regression. Hence the computation does not require special software or packages. We published the R functions of the WEE approach using linear regression, logistic regression, and quantile regression on github (<https://github.com/song-xiaoyu/secondary-analysis-in-case-control-studies>) together with selected codes of comparison methods and simulations. Users can utilize these functions directly for their analyses in R or revise the codes based on their respective needs.

Overall, the WEE provides a very general secondary analysis framework. Through intensive numerical investigations, we found that it is particularly useful and outperforms the existing approaches when $P(D|X, Y)$ do not follow a linear logistic model, the sampling scheme is unknown or complex, or likelihood functions are not available.

Based on our investigations on the secondary analysis in the case-control studies, we provide some suggestions for the selection of methods in a real GWAS study. The first step is investigating the Y and D relationship. If $Y \perp D$, classical methods are valid, and researchers do not need to consider novel approaches to adjust for biases. For studies where Y and D are correlated regardless of the source of correlation, it is worthwhile to consider the sampling scheme. The IPW approach provides robust and similar efficient estimates to those of the WEE approach under a random sample. However, if the sampling scheme is unclear or complex, the WEE could serve as an alternative approach. One may need to be cautious with likelihood-based approaches such as SPML, as the estimates may be biased and the efficiency may not be achieved with a misspecified underlying disease model.

Future extensions: The WEE approach can be extended in a few directions. First, it can be adapted for studies with more complex designs. For example, we can apply the same idea to matched case-control designs and combine their estimating functions of the matched subsets for estimation. The WEE can also be extended to studies where the primary disease includes multiple categories or is continuously sampled but oversampled at the “low” and/or “high” extremes.

Table 8 The estimated allelic effects on log serum IgE level in *quantile regression* at quantile levels of 0.15, 0.25, 0.5, 0.75, and 0.85

SNP	Method	$\tau = 0.15$		$\tau = 0.25$		$\tau = 0.5$		$\tau = 0.75$		$\tau = 0.85$	
		Est.	P-value	Est.	P-value	Est.	P-value	Est.	P-value	Est.	P-value
rs2289276	IPW	-0.1	5.0E-01	-0.1	6.9E-01	0.4	1.1E-02	0.5	6.1E-02	0.1	6.9E-01
	WEE	-0.1	6.4E-02	0.0	9.7E-01	0.4	3.8E-07	0.3	8.7E-04	0.0	7.5E-01
rs1898671	IPW	-0.1	5.4E-01	-0.3	9.5E-02	-0.3	2.1E-01	0.3	3.9E-01	-0.1	7.1E-01
	WEE	-0.2	9.2E-03	-0.2	3.9E-03	-0.2	8.8E-03	0.2	6.7E-02	0.0	9.8E-01
rs11466741	IPW	-0.1	6.5E-01	0.1	5.1E-01	0.5	3.6E-03	0.4	1.5E-02	0.2	3.9E-01
	WEE	0.0	4.1E-01	0.1	1.7E-01	0.4	8.4E-08	0.3	3.5E-04	0.1	3.6E-01
rs11466743	IPW	0.4	6.7E-01	0.0	9.4E-01	-0.5	8.2E-02	-1.1	3.6E-02	-1.1	1.4E-02
	WEE	0.4	1.6E-01	0.0	9.7E-01	-0.6	1.2E-03	-1.2	1.2E-11	-1.1	5.4E-08
rs2289277	IPW	-0.1	5.5E-01	-0.1	7.1E-01	0.3	2.2E-01	0.3	1.3E-01	0.2	4.6E-01
	WEE	-0.1	7.1E-02	0.0	5.5E-01	0.3	1.7E-04	0.3	5.7E-03	0.1	3.7E-01
rs2289278	IPW	-0.1	7.1E-01	-0.3	2.7E-01	-0.1	7.6E-01	-0.5	2.3E-01	-0.5	3.1E-01
	WEE	-0.1	1.6E-01	-0.3	2.1E-03	-0.1	2.2E-01	-0.5	3.4E-04	-0.3	7.1E-02
rs11241090	IPW	0.4	4.7E-01	0.4	3.4E-01	-0.3	5.9E-01	-0.1	9.4E-01	0.4	6.1E-01
	WEE	0.4	3.2E-02	0.3	2.3E-02	-0.3	1.6E-01	0.0	9.3E-01	0.3	2.8E-01
rs10035870	IPW	0.7	1.1E-01	0.5	2.3E-01	0.8	2.6E-01	0.8	2.0E-02	0.4	3.1E-01
	WEE	0.8	1.1E-09	0.6	9.0E-07	0.9	5.4E-07	0.8	3.1E-04	0.4	3.7E-02
rs11466749	IPW	-0.2	3.8E-01	-0.2	5.3E-01	0.1	8.8E-01	0.2	6.9E-01	0.3	5.6E-01
	WEE	-0.2	2.2E-02	-0.2	6.8E-02	0.0	7.1E-01	0.1	2.7E-01	0.3	7.9E-03
rs11466750	IPW	-0.1	5.0E-01	-0.1	6.6E-01	-0.3	8.2E-02	-0.2	5.1E-01	-0.1	8.7E-01
	WEE	-0.2	3.3E-02	-0.1	5.5E-02	-0.3	2.1E-04	-0.3	4.2E-03	0.1	4.2E-01

IPW, inverse probability weighting quantile regression; WEE, the proposed weighted estimating equation approach estimated by generating pseudo counterfactual observations with $T = 10$ replicates.

The conditional primary disease prevalence $f(D|X, Z)$ can be estimated using proportional odds regression or multinomial logistic regression for categorical primary disease and certain parametric likelihoods for continuous primary disease.

Another extension is to consider meta-/mega-analysis, combining WEE estimates from multiple case-control studies. As most of the case-control studies are powered for the primary

analysis, the power for detecting the important SNPs of secondary traits using a single data set can be limited. If multiple case-control studies exist that measure the same secondary trait, combining them could greatly enhance the power. Another way to improve the power is to model multiple correlated secondary traits jointly. If a SNP is associated with multiple traits, analyzing them jointly could improve the detecting power. The WEE approach can also be extended to multivariate

Figure 1 The estimated distribution functions of log serum IgE level associated with SNPs rs10035870 and rs11466743.

regressions by replacing the estimating function S by the generalized estimating functions as in Liang *et al.* (1992).

In summary, the construction of the WEE is straightforward and computationally efficient by evaluating the expected counterfactual estimating functions or generating pseudo observations. It provides a robust and fairly efficient unbiased estimation for the marker–secondary phenotype association for multiple types of secondary traits. It also has appropriate type I error rate and relatively large power and has proved robust against disease prevalence misspecification. Finally, the WEE can be extended to multiple study designs and can be applied to multiple regressions.

Acknowledgments

This research is supported by the National Science Foundation (DMS-120923) and by the National Institutes of Health (1R03HG007443-01). I.I.-L. is supported by National Institute of Mental Health grant MH095797 and by National Science Foundation grant DMS-1100279.

Literature Cited

- Burrows, B., F. D. Martinez, M. Halonen, R. A. Barbee, and M. G. Cline, 1989 Association of asthma with serum IgE levels and skin-test reactivity to allergens. *N. Engl. J. Med.* 320: 271–277.
- Cornelis, M. C., A. Agrawal, J. W. Cole, N. N. Hansel, K. C. Barnes *et al.*, 2010 The gene, environment association studies consortium (GENEVA): maximizing the knowledge obtained from GWAS by collaboration across studies of multiple conditions. *Genet. Epidemiol.* 34: 364–372.
- Flaherman, V., and G. W. Rutherford, 2006 A meta-analysis of the effect of high weight on asthma. *Arch. Dis. Child.* 91: 334–339.
- Frayling, T. M., N. J. Timpson, M. N. Weedon, E. Zeggini, R. M. Freathy *et al.*, 2007 A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science* 316: 889–894.
- Ghosh, A., F. A. Wright, and F. Zou, 2013 Unified analysis of secondary traits in case–control association studies. *J. Am. Stat. Assoc.* 108: 566–576.
- Guo, S., and M. W. Fraser, 2010 Propensity score analysis. *Stat. Methods Appl.* SAGE publication.
- He, J., H. Li, A. C. Edmondson, D. J. Rader, and M. Li, 2012 A Gaussian copula approach for the analysis of secondary phenotypes in case-control genetic association studies. *Biostatistics* 13: 497–508.
- Jiang, Y., A. Scott, and C. Wild, 2006 Secondary analysis of case-control data. *Stat. Med.* 25: 1323–1339.
- Koenker, R., and G. Bassett, Jr., 1978 Regression quantiles. *Econometrica* 46: 33–50.
- Li, H., and M. Gail, 2012 Efficient adaptively weighted analysis of secondary phenotypes in case-control genome-wide association studies. *Hum. Hered.* 73: 159–173.
- Liang, K.-Y., S. L. Zeger, and B. Qaqish, 1992 Multivariate regression analyses for categorical data. *J. R. Stat. Soc. Ser. B* 54: 3–40.
- Lin, D., and D. Zeng, 2009 Proper analysis of secondary phenotype data in case-control association studies. *Genet. Epidemiol.* 33: 256–265.
- Liu, M., L. Rogers, Q. Cheng, Y. Shao, M. E. Fernandez-Beros *et al.*, 2011 Genetic variants of TSLP and asthma in an admixed urban population. *PLoS ONE* 6: e25099.
- Monsees, G., R. Tamimi, and P. Kraft, 2009 Genome-wide association scans for secondary traits using case-control samples. *Genet. Epidemiol.* 33: 717–728.
- Morjaria, J. B., and R. Polosa, 2009 Off-label use of omalizumab in non-asthma conditions: new opportunities. *Expert Rev. Respir. Med.* 3: 299–308.
- Pritchard, J. K., M. Stephens, and P. Donnelly, 2000 Inference of population structure using multilocus genotype data. *Genetics* 155: 945–959.
- Regan, E. A., J. E. Hokanson, J. R. Murphy, B. Make, D. A. Lynch *et al.*, 2010 Genetic epidemiology of COPD (COPDGene) study design. *J. Chron. Obstruct. Pulmon. Dis.* 7: 32–43.
- Richardson, D., P. Rzehak, J. Klenk, and S. Weiland, 2007 Analyses of case-control data for additional outcomes. *Epidemiology* 18: 441–445.
- Van der Vaart, A. W., 2000 *Asymptotic Statistics*, Vol. 3. Cambridge University Press, Cambridge/London/New York.
- Visscher, P. M., M. A. Brown, M. I. McCarthy, and J. Yang, 2012 Five years of GWAS discovery. *Am. J. Hum. Genet.* 90: 7–24.
- Wang, J., and S. Shete, 2011 Estimation of odds ratios of genetic variants for the secondary phenotypes associated with primary diseases. *Genet. Epidemiol.* 35: 190–200.
- Wang, J., and S. Shete, 2012 Analysis of secondary phenotype involving the interactive effect of the secondary phenotype and genetic variants on the primary disease. *Ann. Hum. Genet.* 76: 484–499.
- Wei, J., R. J. Carroll, U. U. Müller, I. V. Keilegom, and N. Chatterjee, 2013 Robust estimation for homoscedastic regression in the secondary analysis of case–control data. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 75: 185–206.
- Wei, Y., A. Pere, R. Koenker, and X. He, 2006 Quantile regression methods for reference growth charts. *Stat. Med.* 25: 1369–1382.
- Wei, Y., X. Song, M. Liu, I. Ionita-Laza, and J. Reibman, 2015 Quantile regression in the secondary analysis of case-control data. *J. Am. Stat. Assoc.* (in press).

Communicating editor: I. Hoeschele

Appendix

The WEE in Survival Analysis

The key for using the new estimating equations is to find the estimating function $S(X, Y, \mathbf{Z}, \beta)$ of each respective outcome and then to plug it into the basis of the estimating equations:

$$E_Y[S(X, Y, \mathbf{Z}, \beta^*)|X, \mathbf{Z}] = E_Y[S(X, Y, \mathbf{Z}, \beta^*)|X, \mathbf{Z}, D = 0]P(D = 0|X, \mathbf{Z}) + E_Y[S(X, Y, \mathbf{Z}, \beta^*)|X, \mathbf{Z}, D = 1]P(D = 1|X, \mathbf{Z}) = 0.$$

For example, we consider a Cox proportion hazards model with the fixed covariates for time-to-event data. Let T_i , C_i , X_i , and $\mathbf{Z}_i$ be the event time, the censoring time, the genetic variant, and the covariate vector of individual i , respectively, for $i = 1, \dots, n$. The observed time of individual i is defined by $Y_i = \min(T_i, C_i)$ and $\delta_i = I(T_i \leq C_i)$. Then, the hazard function of subject i is $\lambda_i(t) = \lambda_0 \exp(\beta_0 + \beta_1 X_i + \beta_2 \mathbf{Z}_i)$. The corresponding partial-likelihood function is

$$L(\beta) = \prod_{i=1}^n \left[\frac{\exp(\beta_0 + \beta_1 X_i + \beta_2 \mathbf{Z}_i)}{\sum_{j \in R(Y_i)} \exp(\beta_0 + \beta_1 X_j + \beta_2 \mathbf{Z}_j)} \right]^{\delta_i},$$

where $R(t) = \{j | Y_j \geq t\}$,

and the partial score function is

$$S(\beta) = \frac{\partial \log(L)}{\partial \beta_1} = \sum_{i=1}^n \delta_i [X_i - \bar{X}_i],$$

where $\bar{X}_i = \sum_{j \in R(Y_i)} w_j X_j$ and $w_j = \exp(\beta_0 + \beta_1 X_j + \beta_2 \mathbf{Z}_j) / \sum_{j \in R(Y_i)} \exp(\beta_0 + \beta_1 X_j + \beta_2 \mathbf{Z}_j)$.

With the partial score function, we then plug it into the basis to construct the new estimating equations and solve it for the unbiased estimation of the association between X and time-to-event secondary trait Y in a case-control study. The new set of estimating equations can be solved by generating pseudo observations for each subject in cases and controls or by estimating expectation of the counterfactual estimating functions, assuming a known weight $P(D|X, \mathbf{Z})$. The weight $P(D|X, \mathbf{Z})$ can be solved by modeling the primary disease through a logistic regression.